U.S. DEPARTMENT OF ENERGY | Office of Science

SciDAC Scientific Discovery through Advanced Computing

Scaling the Earth System Grid to 100 Gbps: Network Architecture Challenges

Eli Dart[1], Alex Sim[1], Brian Tierney[1], and Dean N. Williams[2]

[1]Lawrence Berkeley National Laboratory; [2]Lawrence Livermore National Laboratory

Summary

The climate community faces a rising tide of data as high performance computing enters the petascale era and more sophisticated Earth system models provide greater scientific utility. With many extraordinarily large data warehouses located globally, researchers will depend heavily on high performance networks to access distributed data, information, models, analysis, visualization tools, and other computational resources. In this highly collaborative decentralized problem-solving environment, a faster network – on the order of 10 to 100 times faster than what exists today – will be needed to most efficiently use the data and tools available to scientists. Climate researchers want the ability to combine multiple data sets, which may be up to 300TB each, for analysis. This is not feasible using today's 10 Gigabit per second networks. Therefore we need to ensure the ESG architecture scales to the next generation network speeds of 100 Gbps. The Earth System Grid Center for Enabling Technologies (ESG-CET) project is working closely with the Energy Sciences network (ESnet) to make high-speed data federation a reality.

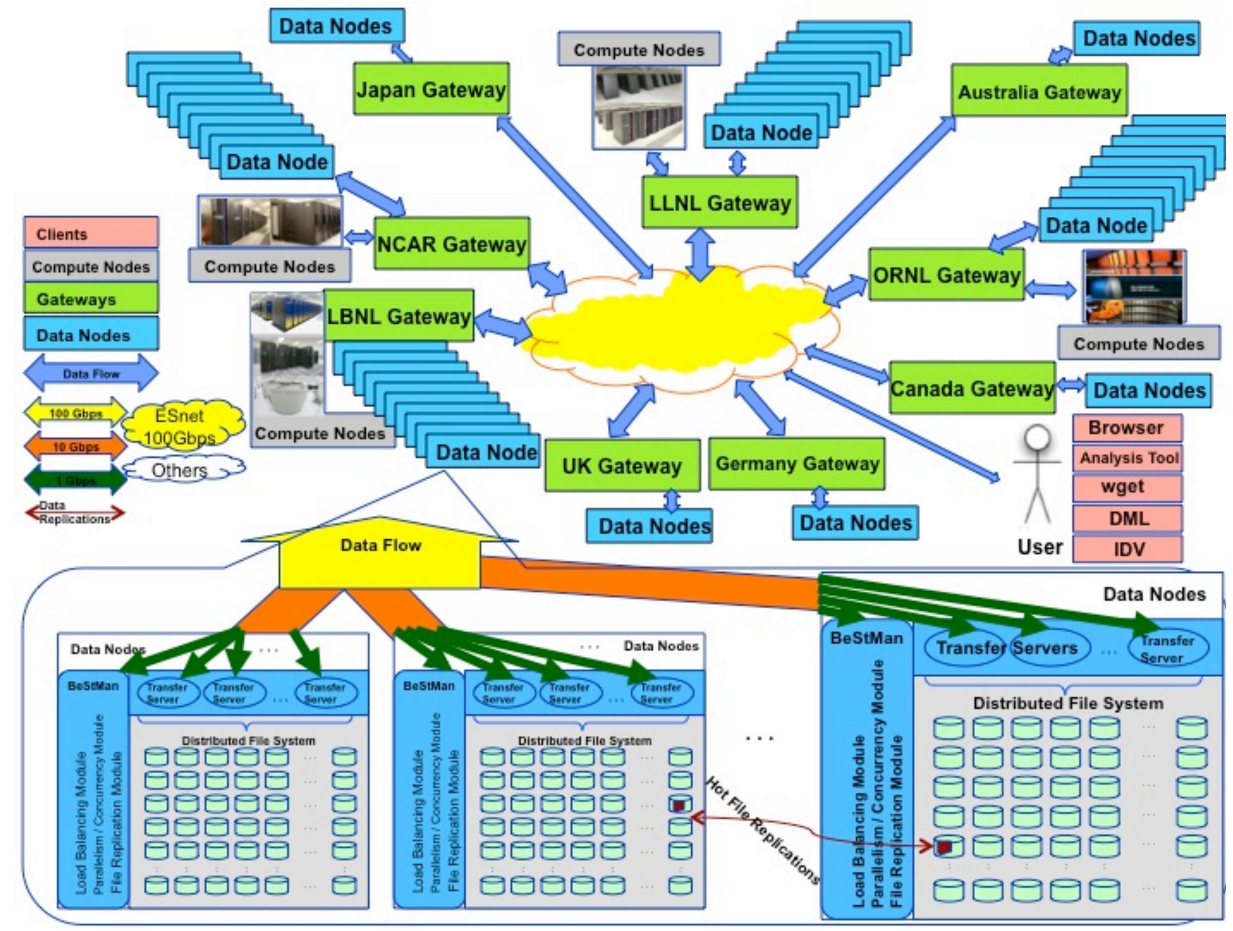

Figure 1. The envisioned federated topology of the ESG-CET enterprise system utilizing 100 Gigabits per second (Gbps) network connections. A network of geographically distributed Gateways and Data Nodes is built into a globally federated "built-to-share" scientific discovery infrastructure. By federating these Gateways using a fast network, independent data warehouses deliver seamless access to vast data archives to scientists and their specialized client applications. Experts (e.g., model developers, climate researchers) and non-experts alike need fault-tolerant end-to-end system integration and large data movement, and benefit from rich data exploration and manipulation – in the process moving vast amounts of data to and from sites around the world.

As ESG-CET takes its first steps towards building a "science gateway" – open to everyone – it must take into consideration the network (and network performance) in relation to distributed data sharing processes associated with data discovery, access, movement, analysis, visualization and computational resources. The Earth System Grid's (ESG) federated architecture needs to be able to scale in order to handle increasing demand for network throughput.

ESnet is moving forward with plans to scale up to 100 Gbps in support of ESG and other DOE science. ESnet's new network, ESnet4, is a next-generation infrastructure built to serve science, and is capable of scaling to 100 Gbps and beyond. Virtual circuit capabilities, such as bandwidth and quality of service guarantees, provided by the ESnet4 Science Data Network (SDN), will allow ESG to manage high volumes of network traffic in a manner that does not impact the network performance of other ESnet users. ESnet's OSCARS service allows users to reserve a virtual circuit of any capacity (from 1 Mbps up to the fastest link in the network, which will soon be 100 Gbps), depending on application requirements.

While compute servers in the ESG architecture will lessen the need for network traffic to end user sites by doing data reduction and data manipulation on the server side, there will still be a need to move vast amounts of data to and from sites for scientific purposes (e.g., model ensembles, model inter-comparisons, etc.). In the ESG architecture, data reduction is done on the front-end server (i.e., Data Node) that is serving the data to the client. This means that each Data Node will need significant compute power and may use a little less network bandwidth when doing data reduction. In the future, many nodes will have to work in parallel, and this will require 100 Gbps ESnet links. (Gateways are where data are queried and located, while Data Nodes accept data that are published to the ESG system, store the data, and serve the data to clients.) We assume any given data node will be capable of data rates around 800 Mbps. When ESG scales to 125 nodes per each site then an entire 100 Gbps ESnet link will be required.

The biggest challenge for utilizing 100 Gbps network speeds in the ESG architecture is in the coordination and management of the file system I/O up to the available maximum bandwidth. Multiple Data Nodes can be used to scale the system as bandwidth requirements increase. These Data Nodes sit in front of a large distributed file system that can move data files in parallel and concurrently. This requires coordinated access to underlying multiple distributed file systems.

The replication and aggregation described here will result from ESG developing a system capable of providing load balancing across a large number of Data Nodes, the ability to replicate data as necessary, and includes a replica location service to forward data requests to the proper Data Node. Berkeley Storage Manager (BeStMan) and the Replica Location Service (RLS) – both ESG software components – provide much of the functionality to do this already. Experience suggests that this will scale to the increasing number of client data requests as well as higher bandwidth. We know that the High Energy Physics community has been putting efforts in such a system, and an experiment group in University of California at San Diego, for example, has 110 data transfer servers, which will grow to about 200 data transfer servers in the near future, resulting in very high data transfer rates. ESG will do something similar for each data node site.

The complexity, sophistication, and rigor of the climate models must continue to increase dramatically to get the precision needed to predict climate change. The climate modeling community of tomorrow will have a tremendous need for the ESG-CET (or its descendents), and for the network backbone to be fully functional, lightweight, fast, and accurate enough to study climate change.

For further information on this subject contact:
Name: Dean N. Williams.
Organization: Lawrence Livermore National Laboratory
Email: williams13@llnl.gov
Phone: (925) 423-0145